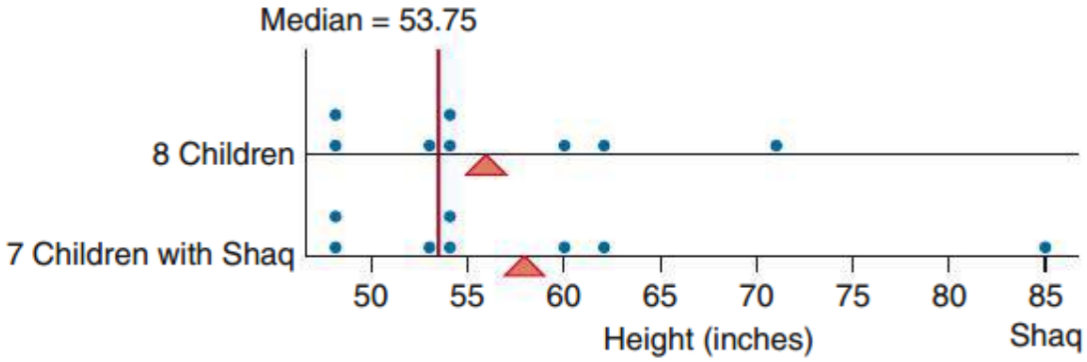
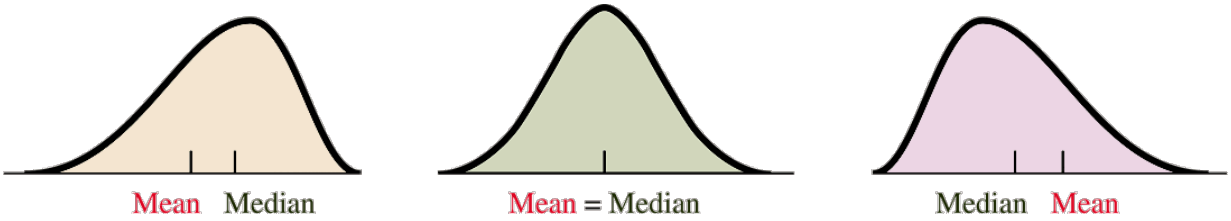
3.4: Comparing Measures of Center

Shape	Measure for center	Measure for spread
Symmetric	Mean	Standard deviation
Skewed	Median	IQR

- Skewed data and outliers affect the mean and standard deviation
- The median is resistant to outliers; it is not affected by the size of an outlier

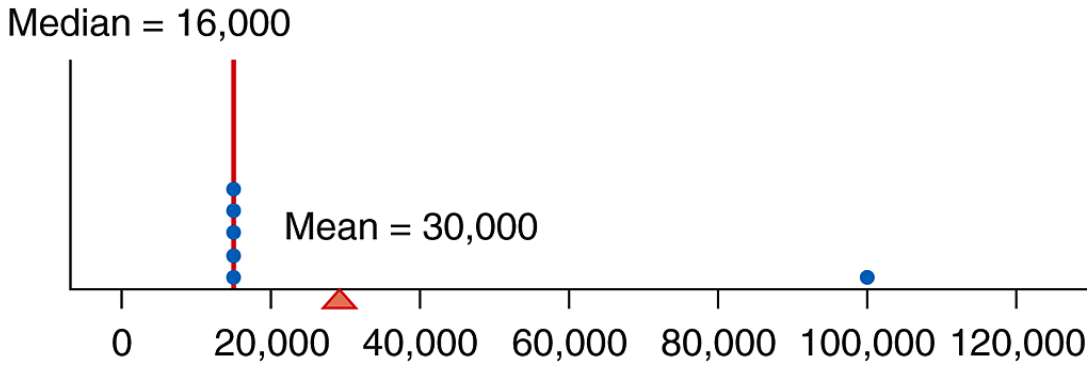


Shape	Mean vs. Median
Skewed left	Mean $<$ Median
Symmetric	Mean $=$ Median
Skewed right	Mean $>$ Median



Example. A (very small) fast-food restaurant has five employees, all of whom work full-time for $7 per hour. Each employee's annual income is about $16,000 per year. The owner, on the other hand, makes $100,000 per year.

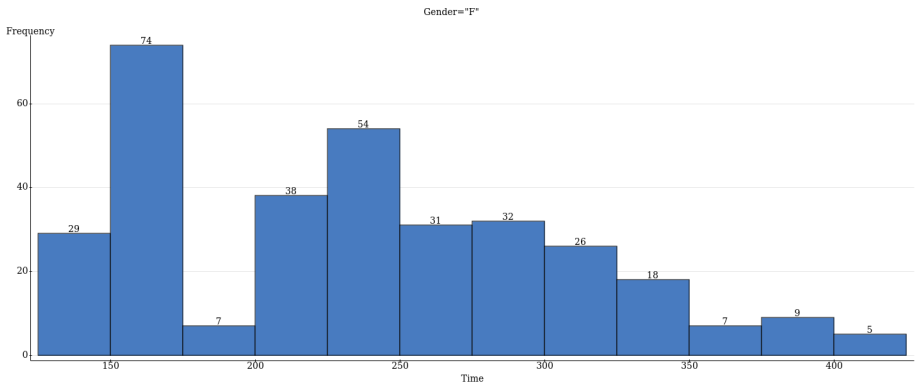
Find both the mean and the median. Which would you use to represent the typical income at this business – the mean or the median? Which value is smaller?



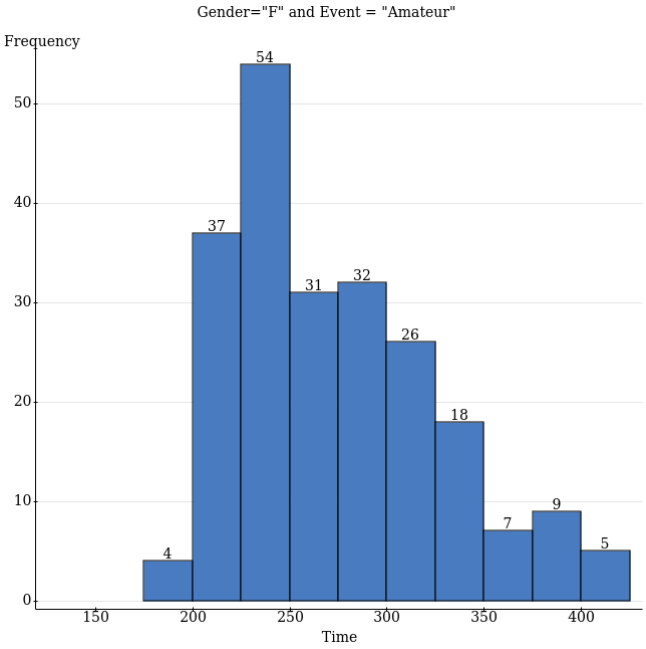
When comparing distributions:

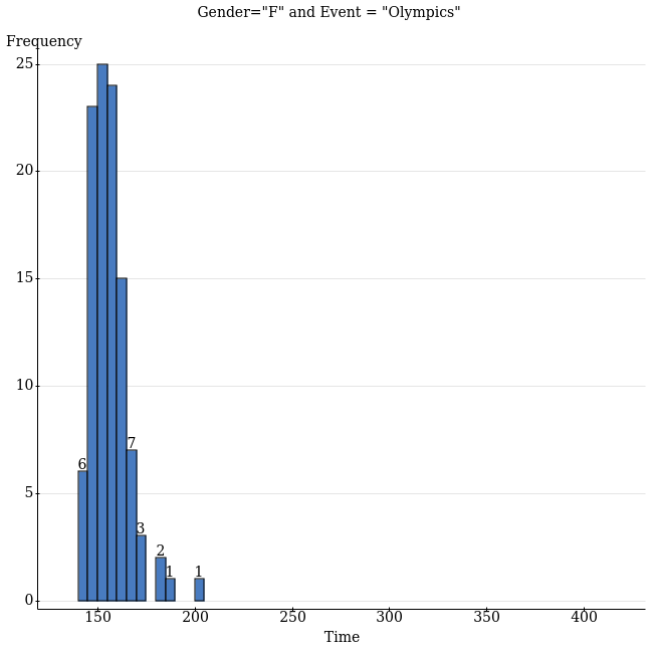
- Always use the same measures of center and spread for both distributions.
- If one of the distributions is skewed, use Median and IQR to compare both!

Example. Below is a histogram of the finishing times of female marathon runners.



If we separate the data into the "Amateur" and "Olympic" events, we see why the data is bimodal. If we compare the distributions, should we use the Mean or the Median? Should we use the standard deviation, or the IQR?

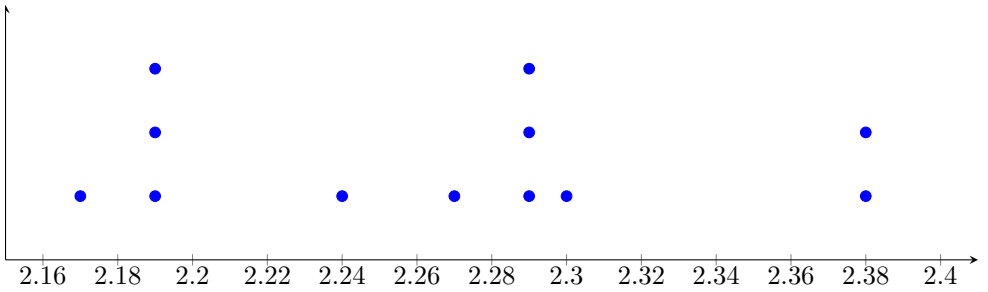


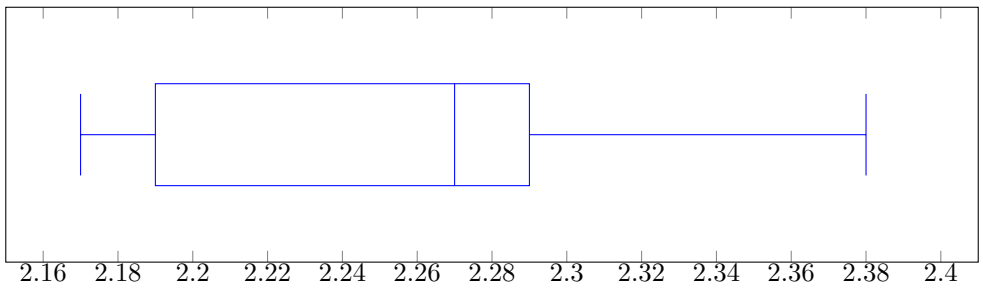


3.5: Using Boxplots for Displaying Summaries

Definition.
A **boxplot** is a graphical tool for visualizing a distribution. Boxplots can be useful for comparing multiple distributions. In a box plot:

- The left edge of the box represents Q_1
- The vertical line inside the box represents the median (Q_2)
- The right edge of the box represents Q_3
- Lines extending past the edges of the box are called whiskers. The whiskers extend to the most extreme values that are not *potential* outliers.



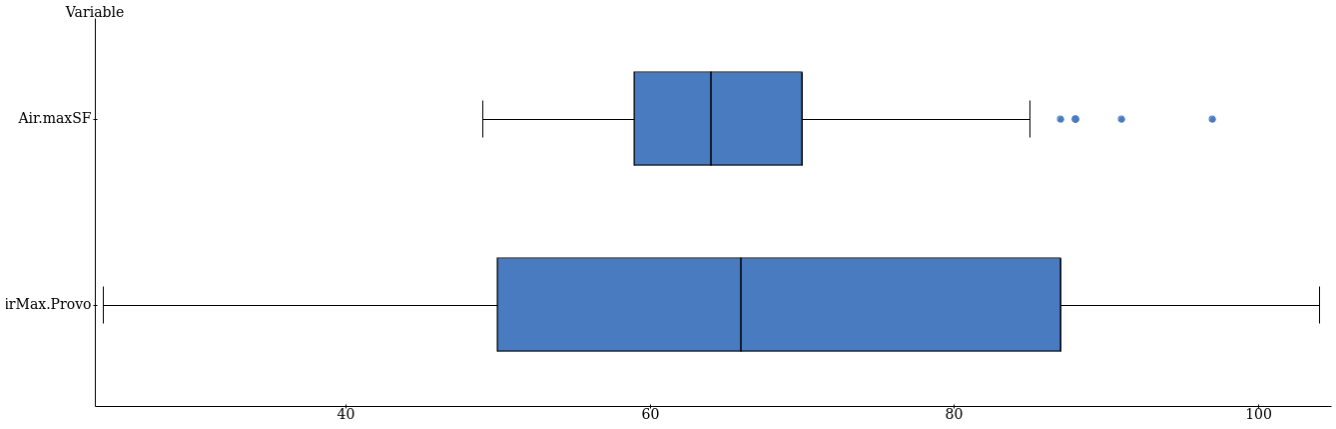


Definition.
Potential outliers are any values that are

- less than $Q_1 - 1.5IQR$
- more than $Q_3 + 1.5IQR$

These values are the left and right limits. They are also known as the *fences*.

Example. Using the "airtemp" dataset in StatCrunch, generate the boxplots for the daily maximum temperature in San Francisco and Provo. Compute the left and right limits. Are they included in the plots?



Definition.
The **five number summary** is

the minimum, Q_1, the median, Q_3, the maximum